■ Fiche de dépôt / Deposit Record – Corpus linguistique breton Nidiad (collection complète)

1. Identification générale / General Identification

FR :
- Titre complet : Nidiad — Corpus linguistique breton (collection complète)
- Format : JSON Lines (UTF-8)
- Langue principale : breton (br)
- Nombre de volumes : variable selon la série éditoriale
- Type de contenu : textes littéraires, scientifiques, pédagogiques et lexicographiques
- Type de données : texte brut nettoyé, segmenté par page (1 paragraphe long par page)
- Périmètre : ensemble des volumes publiés par Preder & Nidiad

EN :
- Full title: Nidiad — Breton Linguistic Corpus (complete collection)
- Format: JSON Lines (UTF-8)
- Main language: Breton (br)
- Number of volumes: variable, according to editorial series
- Content type: literary, scientific, pedagogical, and lexicographic texts
- Data type: cleaned raw text, segmented by page (1 long paragraph per page)
- Scope: all volumes published by Preder & Nidiad

2. Origine et édition / Source and Edition

FR :
- Éditeur scientifique : Preder
- Collection : Nidiad
- Numérisation et nettoyage : 2025 — traitement linguistique approfondi automatisé (extraction, recomposition et structuration pour usage IA)
- Responsable du traitement : Preder / Nidiad
- Années couvertes : 2000–2025
- Mode de production : extraction à partir de fichiers PDF ou sources typographiques, conversion en texte brut, normalisation orthographique et diacritique

EN :
- Scientific editor: Preder
- Collection: Nidiad
- Digitization and cleaning: 2025 — advanced automated linguistic processing (extraction, reconstruction, and structuring for AI use)
- Processing supervisor: Preder / Nidiad
- Coverage years: 2000–2025
- Production method: extraction from PDF or typographic sources, conversion to plain text, orthographic and diacritic normalization

3. Licence et droits / License and Rights

FR :
- Licence : CC-BY 4.0 Preder & Nidiad (selon Lañvaz Preder)
- Conditions d’usage :
• Autorisé : entraînement et évaluation d’IA de traduction, d’analyse linguistique et d’enseignement.

• Interdit : modification du vocabulaire, de l’orthographe ou de la grammaire du breton.
• Mention obligatoire : Source : Preder & Nidiad, corpus breton Nidiad (CC-BY 4.0 — usage IA autorisé, intégrité linguistique garantie).

EN :
- License: CC-BY 4.0 Preder & Nidiad (as defined in Lañvaz Preder)
- Usage terms:
• Permitted: AI training and evaluation for translation, linguistic analysis, and education.
• Prohibited: modifying Breton vocabulary, spelling, or grammar.
• Required attribution: Source: Preder & Nidiad, Nidiad Breton corpus (CC-BY 4.0 — AI use authorized, linguistic integrity guaranteed).

4. Description linguistique et méthode / Linguistic Description and Method

FR :
- Objectif linguistique : constituer un corpus de référence en breton moderne normalisé.
- Structure interne : un fichier JSONL par volume, un paragraphe long par page.
- Méthode de traitement :
• extraction ligne par ligne depuis les fichiers PDF ;
• suppression des artefacts (en-têtes, numéros, crédits, coupures) ;
• recomposition automatique des mots segmentés (mot-\nsuite → mots complets) ;
• respect intégral des diacritiques bretons et de la segmentation des phrases ;
• encodage UTF-8 et vérification page par page.

EN :
- Linguistic goal: to create a reference corpus for standardized modern Breton.
- Internal structure: one JSONL file per volume, one long paragraph per page.
- Processing method:
• line-by-line extraction from PDF files;
• removal of layout artifacts (headers, numbering, credits, splits);
• automatic reconstruction of hyphenated words (word-\npart → full word);
• full preservation of Breton diacritics and sentence segmentation;
• UTF-8 encoding and page-by-page verification.

5. Objectif du corpus / Corpus Purpose

FR :
- Constituer un ensemble linguistique homogène destiné à :
• entraînement de modèles de traduction breton ↔ français ;
• amélioration de la reconnaissance linguistique automatique ;
• apprentissage grammatical et lexicographique ;
• enrichissement du lexique grammatical de la collection Preder / Nidiad.

EN :
- To build a homogeneous linguistic dataset for:
• training Breton ↔ French translation models;
• improving automatic linguistic recognition;
• grammatical and lexicographic learning;
• enriching the grammatical lexicon of the Preder / Nidiad collection.

6. Vérification technique / Technical Verification

FR :
- Encodage : UTF-8 validé ■
- Structure : JSONL (un paragraphe long par page) ■
- Cohérence linguistique : orthographe, diacritiques et segmentation vérifiés ■
- Interopérabilité : conforme aux formats IA standard (JSON, JSONL, UTF-8)

EN :
- Encoding: UTF-8 validated ■
- Structure: JSONL (one long paragraph per page) ■
- Linguistic consistency: orthography, diacritics, and segmentation verified ■
- Interoperability: compliant with standard AI formats (JSON, JSONL, UTF-8)